DOI: 10.1534/genetics.104.033910

Model Selection in Binary Trait Locus Mapping

Cynthia J. Coffman,*,† R. W. Doerge,‡,§ Katy L. Simonsen,‡ Krista M. Nichols, Christine K. Duarte,††,‡‡ Russell D. Wolfinger‡‡ and Lauren M. McIntyre†,§,§§,1**

**Institute for Clinical and Epidemiological Research Biostatistics Unit, Durham VA Medical Center (152), Durham, North Carolina 27705, †Department of Biostatistics and Bioinformatics, Duke University Medical Center, Durham, North Carolina 27710, ‡Department of Statistics, Purdue University, West Lafayette, Indiana 47907-2068, §Department of Agronomy, Purdue University, West Lafayette, Indiana 47907, §§Computational Genomics, Purdue University, West Lafayette, Indiana 47907, **Washington State University, School of Biological Sciences, Pullman, Washington 99164, ††Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695 and ‡‡SAS Institute, Cary, North Carolina 27513*

Manuscript received July 23, 2004
Accepted for publication March 17, 2005

ABSTRACT

Quantitative trait locus (QTL) mapping methodology for continuous normally distributed traits is the subject of much attention in the literature. Binary trait locus (BTL) mapping in experimental populations has received much less attention. A binary trait by definition has only two possible values, and the penetrance parameter is restricted to values between zero and one. Due to this restriction, the infinitesimal model appears to come into play even when only a few loci are involved, making selection of an appropriate genetic model in BTL mapping challenging. We present a probability model for an arbitrary number of BTL and demonstrate that, given adequate sample sizes, the power for detecting loci is high under a wide range of genetic models, including most epistatic models. A novel model selection strategy based upon the underlying genetic map is employed for choosing the genetic model. We propose selecting the "best" marker from each linkage group, regardless of significance. This reduces the model space so that an efficient search for epistatic loci can be conducted without invoking stepwise model selection. This procedure can identify unlinked epistatic BTL, demonstrated by our simulations and the reanalysis of *Oncorhynchus mykiss* experimental data.

STATISTICAL methods for mapping single genes for continuous and binary traits in experimental populations have advanced significantly in the past few years (Lander and Botstein 1989; Haley and Knott 1992; Zeng 1994; Satagopan *et al.* 1996; Xu 1996; Xu and Atchley 1996; Yi and Xu 2000; McIntyre *et al.* 2001; Yi and Xu 2002). Single-gene QTL models have been expanded to encompass multiple-QTL mapping problems by using cofactors or additional markers (Jansen 1993; Zeng 1994; Xu 2003). Multiple-QTL models have been developed for both continuous and binary traits (Sillanpää and Arjas 1998; Kao *et al.* 1999; Zeng *et al.* 2000; Jannink and Jansen 2001; Sen and Churchill 2001; Carlborg and Andersson 2002; Yi and Xu 2002; Xu 2003) and for discrete traits with multiple observation classes (Yi *et al.* 2004). When attempting to identify multiple QTL for a trait, model selection is a key issue as the number of possible models quickly becomes large. In most analyses, the enumeration of all QTL models for a data set is possible only when the number of markers is limited. An exception is a recent method (Xu 2003) that estimates the effect for all markers, thus avoiding the testing and model selection issues.

One approach for reducing the dimensionality of the model space is to locate all QTL that are significantly associated with the trait, using single-QTL methods, and then build the multiple-QTL models using only the QTL selected in the single-gene analysis (Kao *et al.* 1999). When all QTL are additive, single-marker analysis is a reasonable strategy for identifying QTL (Coffman *et al.* 2003). However, epistasis can alter the trait in a manner that may be difficult to predict (Doerge 2001), thus further complicating the model fitting and selection process. Carlborg *et al.* (2000) proposed a method for simultaneous mapping of pairwise interacting QTL. In addition, Carlborg and Andersson (2002) proposed a forward selection strategy that incorporates a randomization test to identify epistatic QTL. Unfortunately, this approach will miss pairs of loci that are epistatic without a contributing main effect. Holland *et al.* (2002) performed a pairwise grid search to identify potential epistatic loci and then include the most significant pairs in the "best" single-gene model via a forward stepwise procedure. Yi and Xu (2002) proposed a Bayesian method to map multiple QTL with pairwise locus epistasis. Sen and Churchill (2001) also presented a Bayes-

[1]*Corresponding author:* Department of Agronomy, 915 W. State St., Purdue University, West Lafayette, IN 47907.
E-mail: lmcintyre@purdue.edu

ian analysis that implements a strategy similar to that of Jansen (1993), where the QTL problem is divided into two pieces, detection and then localization. While all of these approaches have a common goal, the complexity and computational intensity of many of these approaches make them difficult to implement. Furthermore, stepwise procedures and pairwise searches do not investigate the entire model space and these approaches have been shown to fail to identify all possible effects in different applications (Harrell 2001; Burnham and Anderson 2002).

Searching through the potential models to identify the best model is an active area of statistical research (Harrell 2001; Burnham and Anderson 2002). Several common criteria are used to judge and compare models to select the best model. Due to the large number of models that may be examined in these analyses, issues of model selection bias and uncertainty should be addressed (Burnham and Anderson 2002). In the method described by Jansen (1993), the Akaike information criterion (AIC) was used for model selection. Broman and Speed (2002) reviewed different model selection criteria for QTL analysis and proposed a criterion that is a modification of the Bayesian information criterion (BIC) (Schwarz 1978). Sillanpää and Corander (2002) gave a general review of model selection criteria and advocated the Bayesian idea of model averaging. Others are working on modifications of these criteria to improve their performance in the QTL setting (Ball 2001; Bogdan *et al.* 2004; Siegmund 2004). However, these criteria have not been specifically evaluated for binary traits.

In genetic experiments, binary traits often occur when considering characteristics related to susceptibility/resistance, sterility/fertility, and mortality/survival. Sen and Churchill (2001) examined binary traits using a generalized linear model framework. Yi and Xu (2000) proposed a Bayesian method for complex binary traits under the threshold model and later extended this method to map multiple QTL with pairwise locus epistasis for binary traits (Yi and Xu 2002). Kilpikari and Sillanpää (2003) present a multilocus Bayesian approach for association mapping that can be used for binary traits under the threshold or liability model. The threshold model is an important quantitative genetic model. However, the underlying threshold distribution is unobserved (Falconer and Mackay 1996; Lynch and Walsh 1998), presenting challenges in specifying the functional form of the threshold model.

In the human genetics literature, binary traits (disease status) are often parameterized in terms of the penetrance as well as the physical distance. This model formation has been routinely employed for segregation and linkage analysis (Ott 1991; Gauderman and Thomas 2001). The value of the penetrance parameter can be estimated as a part of segregation analysis or in a joint segregation and linkage analysis. The concept of incomplete penetrance is important, as it underscores the complexity encountered in analysis of binary traits.

As an adaptation of the model in human genetics, and extension of previous work in experimental populations (McIntyre *et al.* 2001), we propose a method to detect and estimate multiple binary trait loci (BTL). We focus on the case where penetrance is incomplete and the population structure is a backcross or F_2 from two inbred parents. Using the biological information in the linkage groups, the model space is reduced by choosing the best marker in each linkage group. Consequently, all possible models can be enumerated and stepwise selection procedures are avoided, which in turn eliminates the need for computationally intensive model space exploration. We use a general probability model based on classical transmission genetics to develop a likelihood for the binary phenotype (Simonsen 2004) to estimate recombination and penetrance for multiple BTL under complex genetic models for an experimental population. Regression models are fitted on the basis of this likelihood (Haley and Knott 1992; Jansen 1992; Jansen and Stam 1994; Whittaker *et al.* 1996; Thompson 1998), using a cell means model parameterization rather than the factor effects parameterization (Kutner *et al.* 2004). The parameterization in terms of the cell means clarifies the identification of epistatic loci.

Using simulated data, AIC (Akaike 1973) and BIC (Schwarz 1978) model selection criteria are employed and compared for a limited number of markers as well as in the context of a genome scan. A new SAS procedure, PROC BTL, has been developed and is freely available by request (http:www.genomics.purdue.edu/services/software/btl). PROC BTL includes model selection for a wide range of model selection criteria and implements all of the standard model selection techniques including the one proposed here. Using PROC BTL, we present a reanalysis of *Oncorhynchus mykiss* (rainbow or steelhead trout) data where single-marker associations of the binary trait, resistance to *Ceratomyxa shasta* (a myxozoan parasite) (Nichols *et al.* 2003), suggest that multiple loci may be associated with the resistance. We also find evidence for epistatic effects.

METHODS

A probability model: We denote individual genetic markers by M_i and BTL by G_i, where i indicates the BTL in map order. We assume a map based on k BTL and k markers ($M_1 G_1 M_2 G_2 \cdots M_k G_k$). The complete genotype for all loci is denoted **M** for markers and **G** for BTL, and the possible values are described below.

For a backcross (BC) or F_2 population of diploid individuals from a single cross of homozygote inbred parents, there are only two possible alleles for each marker and/or BTL (denoted by either 1 or 2). In a BC popula-

tion with k loci, there are 2^k distinct marker classes (*i.e.*, possible values for **M**) and 2^k distinct BTL genotypes (*i.e.*, possible values for **G**), giving a total of 4^k possible combinations of genotypic marker classes (**M**, **G**). In an F_2 population with phase unknown, there are 3^k distinct marker classes for **M** and 3^k distinct BTL genotypes for **G**, resulting in a total of 9^k possible combinations of marker classes and genotypes for (**M**, **G**). The number of distinct marker classes or genotypes is represented by K from this point forward (*i.e.*, $K = 2^k$ for BC and $K = 3^k$ for F_2).

As in Simonsen (2004), we label and order the K possible genotypes of markers or BTL as if the genotypes were numerals with one digit per locus, in ascending order. The digit 1 represents the homozygote 1/1 genotype at that locus, 2 is the 1/2 or 2/1 heterozygote, and 3 is the 2/2 homozygote. A genotype for k loci is then a k digit number. Thus, with $k = 2$, a backcross has $K = 4$ possible types {11, 12, 21, 22}, whereas an F_2 has $K = 9$ possible genotypes {11, 12, 13, 21, 22, 23, 31, 32, 33}. We label these K values $m_1, \ldots, m_K$ or $g_1, \ldots, g_K$, depending on whether they represent markers or BTL, respectively. The probability distribution of **M** or **G** specifies the probabilities of each of these K values and thus can be written in a vector of length K in the order given. The joint probability distribution of (**M**, **G**) can be written as a $K \times K$ matrix **Pr**(**M**, **G**), where the rows index the marker classes and the columns index the BTL genotypes. The (i, j)th entry of this matrix represents $\Pr(\mathbf{M} = m_i, \mathbf{G} = g_j)$, where m_i and g_j each take on the K possibilities described above. All matrices and vectors referring to genotypes assume this ordering and indexing.

The recombination rate, r_i, is the probability that an exchange of genetic material (crossover) occurs between the BTL G_i and the marker M_i, where i ranges from 1 to k, where $r_i = 0$ indicates complete association and $r_i = 0.50$ indicates no association between the marker and the BTL. Similarly, the rate of recombination between markers, θ_i, is the probability that an exchange of genetic material occurs between marker M_i and M_{i+1}, where i ranges from 1 to $k - 1$. If the marker map is assumed known, then the θ_i are fixed.

The probability of observing the binary trait is specified by K penetrance parameters, p_j, which are Bernoulli probabilities representing the probability that a binary trait Y is present given a specific BTL genotype j (McIntyre *et al.* 2001). The vector **p** with entries $p_j = \Pr(Y = 1|\mathbf{G} = g_j)$ is of length K, and its jth entry, p_j, is the penetrance parameter for the jth genotype, g_j, where j indexes the possible genotypes in the order explained above. To emphasize the relationship between the genotype and the penetrance the notation p_{g_j} may be used as well as the above p_j. Including a penetrance parameter for each genotype is convenient for visualizing the impact of various genetic models on the parameter space and is a common tool in human genetics (Ott 1991; Gauderman and Thomas 2001).

Trait values can be modeled in a variety of ways, and it is useful to consider the penetrance **p** in the context of standard ANOVA models. For example, consider a backcross with $k = 2$. The penetrance parameters are p_{g_j}, where $g_j = \{11, 12, 21, 22\}$. Suppose the first digit of g_j is s and the second digit is t. The factor effects parameterization would be $p_{g_j} = \mu + \alpha_s + \beta_t + (\alpha\beta)_{st}$, where α and β are the main effects at the two loci and $(\alpha\beta)$ represents the interaction or epistatic effect and $0 \le p_{g_j}, \mu \le 1$. The corresponding cell means model parameterization is $p_{g_j} = \mu_{st}$, where the μ_{st} are (a function of) the cell means. These two model parameterizations are equivalent (Kutner *et al.* 2004). In the factor effects parameterization, absence of epistasis is indicated by $(\alpha\beta)_{st} = 0$ and is equivalent to the constraint in the cell means model of $p_{11} - p_{12} = p_{21} - p_{22}$ (see Figure 2a). If only locus 1 contributes to the trait, the factor effects model constraint is $\beta_t = (\alpha\beta)_{st} = 0$ while the cell means model constraint is $p_{11} = p_{12}$ and $p_{21} = p_{22}$. Similarly, if only locus 2 is involved, the factor effects model constraint is $\alpha_s = (\alpha\beta)_{st} = 0$ and the cell means model constraint is $p_{11} = p_{21}$ and $p_{12} = p_{22}$. Epistasis can be presented as a modification of the expected segregation ratio with fewer than expected phenotypic classes observed (Hartl and Jones 2001). Therefore, the cell means model provides a convenient way of thinking about genetic models, as epistasis is easily defined as equivalence among penetrance parameters (see Figures 2, b–d, and 3).

Simonsen (2004) details the methods for generating the probability model for k BTL in matrix form. The joint probabilities of the BTL genotypes (**G**), marker types (**M**), and the trait (Y), denoted **Pr**(Y, **M**, **G**), can be expressed in terms of **r**, θ, and **p** and generated for k BTL for a specified experimental design. Standard assumptions such as no selection, interference, or mutation are made. As an example, the joint probability distribution of a BC for $k = 2$ is shown in Table 1. The joint probability of every combination of marker and BTL genotype, **Pr**(**M**, **G**), is computed using the recombination probabilities θ and **r**. The matrix for the joint probability of trait, marker, and BTL is then computed by matrix multiplication **Pr**(Y, **M**, **G**) = **Pr**(**M**, **G**) × **Diag**(**p**), since

$$\begin{aligned}\Pr(Y = 1, \mathbf{M} = m_i, \mathbf{G} = g_j) &= \Pr(Y = 1|\mathbf{G} = j)\Pr(\mathbf{M} = m_i, \mathbf{G} = g_j)\\ &= [\mathbf{Pr}(\mathbf{M}, \mathbf{G})]_{i,j} \times p_j.\end{aligned}$$

The joint probability of traits and markers only is used for likelihood calculations as described in the next section. This vector of probabilities is computed as **Pr**(Y, **M**) = **Pr**(**M**, **G**) × **p**, where the matrix multiplication accomplishes the necessary sum over possible genotypes. Its ith entry is

$$\Pr(Y = 1, M = m_i) = \sum_{j=1}^{K} \Pr(Y = 1, M = m_i, G = g_j)$$

TABLE 1

Expected trait distributions for binary traits in a backcross with two markers and two loci for linkage map $M_1G_1M_2G_2$

$M = M_1M_2$	$\Pr(M = m_i)$	$G = G_1G_2$	$\Pr(Y = 1, \mathbf{M} = m_i, \mathbf{G} = g_j)$
$m_1 = 11$	$\frac{1}{2}(1 - \theta_1)$	$g_1 = 11$	$\frac{1}{2}\frac{(1 - r_1)(1 - \theta_1 - r_1)(1 - r_2)}{(1 - 2r_1)}p_{11}$
		$g_2 = 12$	$\frac{1}{2}\frac{(1 - r_1)(1 - \theta_1 - r_1)r_2}{1 - 2r_1}p_{12}$
		$g_3 = 21$	$\frac{1}{2}\frac{r_1(\theta_1 - r_1)(1 - r_2)}{1 - 2r_1}p_{21}$
		$g_4 = 22$	$\frac{1}{2}\frac{r_1(\theta_1 - r_1)r_2}{1 - 2r_1}p_{22}$
$m_2 = 12$	$\frac{1}{2}\theta_1$	$g_1 = 11$	$\frac{1}{2}\frac{(1 - r_1)(\theta_1 - r_1)r_2}{1 - 2r_1}p_{11}$
		$g_2 = 12$	$\frac{1}{2}\frac{(1 - r_1)(\theta_1 - r_1)(1 - r_2)}{(1 - 2r_1)}p_{12}$
		$g_3 = 21$	$\frac{1}{2}\frac{r_1(1 - \theta_1 - r_1)r_2}{(1 - 2r_1)}p_{21}$
		$g_4 = 22$	$\frac{1}{2}\frac{r_1(1 - \theta_1 - r_1)(1 - r_2)}{1 - 2r_1}p_{22}$
$m_3 = 21$	$\frac{1}{2}\theta_1$	$g_1 = 11$	$\frac{1}{2}\frac{r_1(1 - \theta_1 - r_1)(1 - r_2)}{1 - 2r_1}p_{11}$
		$g_2 = 12$	$\frac{1}{2}\frac{r_1(1 - \theta_1 - r_1)r_2}{1 - 2r_1}p_{12}$
		$g_3 = 21$	$\frac{1}{2}\frac{(1 - r_1)(\theta_1 - r_1)(1 - r_2)}{1 - 2r_1}p_{21}$
		$g_4 = 22$	$\frac{1}{2}\frac{(1 - r_1)(\theta_1 - r_1)r_2}{1 - 2r_1}p_{22}$
$m_4 = 22$	$\frac{1}{2}(1 - \theta_1)$	$g_1 = 11$	$\frac{1}{2}\frac{r_1(\theta_1 - r_1)r_2}{1 - 2r_1}p_{11}$
		$g_2 = 12$	$\frac{1}{2}\frac{r_1(\theta_1 - r_1)(1 - r_2)}{1 - 2r_1}p_{12}$
		$g_3 = 21$	$\frac{1}{2}\frac{(1 - r_1)(1 - \theta_1 - r_1)r_2}{1 - 2r_1}p_{21}$
		$g_4 = 22$	$\frac{1}{2}\frac{(1 - r_1)(1 - \theta_1 - r_1)(1 - r_2)}{1 - 2r_1}p_{22}$

Recombination between markers M and BTL G is denoted by r and recombination between markers is denoted by θ. Penetrances are $p_{g_j} = \Pr(Y = 1|\mathbf{G} = g_j)$. Genotypes of the fixed (parental) haplotypes are omitted.

$$= \sum_{j=1}^{K} \Pr(Y = 1|G = g_j)\Pr(M = m_i, G = g_j)$$
$$= \sum_{j=1}^{K} [\mathbf{Pr}(\mathbf{M}, \mathbf{G})]_{i,j} \times p_j.$$

Although the focus of this work is on backcross and F_2 populations, the matrix $\mathbf{Pr}(\mathbf{M}, \mathbf{G})$ can be obtained for any mating scheme. The probability distribution for a generation of offspring can be calculated from the probability distributions for the parental generation, through appropriate matrix operations. By repeating this process any scheme can be derived back to known initial parental generations.

For a $k = 2$ BTL BC population, the four possible (nonfixed) marker allele combinations are $m_1 = 11$, $m_2 = 12$, $m_3 = 21$, $m_4 = 22$, and the matrix rows are given in that order (see Table 1); columns index BTL genotypes in a similar order. Thus row 1 in the matrix $\mathbf{Pr}(Y, \mathbf{M}, \mathbf{G})$ is

$$\frac{1}{2(1 - 2r_1)}[(1 - r_1)(1 - \theta_1 - r_1)(1 - r_2)p_{11} \quad (1 - r_1)(1 - \theta_1 - r_1)r_2p_{12}$$
$$r_1(\theta_1 - r_1)(1 - r_2)p_{21} \quad r_1(\theta_1 - r_1)r_2p_{22}].$$

Likelihood: Using the notation above, the likelihood for observed data $\mathbf{Y} = \mathbf{y}$ and $\mathbf{M} = \mathbf{m}$ is

$$L(\mathbf{r}, \theta, \mathbf{p}) = \Pr(\mathbf{Y} = \mathbf{y}, \mathbf{M} = \mathbf{m}|\mathbf{r}, \theta, \mathbf{p}).$$

This likelihood can also be written in terms of the marker class means, as follows.

The expected marker class means are denoted by the vector $\boldsymbol{\pi}$ whose ith entry, π_i, is the marker class mean for marker class i, namely

$$\pi_i = \mathbf{Pr}(Y = 1|M = m_i) = \frac{\mathbf{Pr}(Y = 1, M = m_i)}{\mathbf{Pr}(M = m_i)}$$

or simply

$$\mathbf{Pr}(Y = 1, M = m_i) = \pi_i \mathbf{Pr}(M = m_i).$$

The component of the likelihood for a single observation with $Y = 1$ and $\mathbf{M} = m_i$ is the ith entry of the vector $\mathbf{Pr}(Y = 1, \mathbf{M})$. Since $\mathbf{Pr}(Y = 0|\mathbf{M} = m_i) = 1 - \pi_i$, we have $\mathbf{Pr}(Y = 0, \mathbf{M} = m_i) = (1 - \pi_i)\mathbf{Pr}(\mathbf{M} = m_i)$. Note that $\boldsymbol{\pi}$ is a function of $\mathbf{r}$, θ, and $\mathbf{p}$, while $\mathbf{Pr}(M)$ is a function of θ only. Therefore, the likelihood for a single observation from marker class i is

$$L(r, \theta, p) = \begin{cases} \pi_i \mathbf{Pr}(M = m_i), & Y = 1 \\ (1 - \pi_i)\mathbf{Pr}(M = m_i), & Y = 0. \end{cases}$$

Suppose in a given sample there are n_i individuals in marker class i, of whom z_i exhibit $Y = 1$, and $n_i - z_i$ exhibit $Y = 0$. Then the likelihood can be written as a product over marker classes:

$$L(\mathbf{r}, \theta, \mathbf{p}) = \prod_{i=1}^{K} (\pi_i \mathbf{Pr}(M = m_i))^{z_i} ((1 - \pi_i)\mathbf{Pr}(M = m_i))^{n_i - z_i} \quad (1)$$

$$= \prod_{i=1}^{K} \pi_i^{z_i} (1 - \pi_i)^{n_i - z_i} (\mathbf{Pr}(M = m_i))^{n_i}. \quad (2)$$

Maximum-likelihood estimates for marker class means: Using this likelihood, the maximum-likelihood estimates (MLEs) for the marker class means can be obtained by maximizing model 1 to obtain estimates of the binomial proportions $\hat{\pi}_i = z_i/n_i$ for $i = 1 \ldots K$.

The marker class means π_i are easily estimated from the data. To estimate penetrance $\mathbf{p}$ and recombination $\mathbf{r}$, we exploit the relationship between $\mathbf{p}$ and $\boldsymbol{\pi}$. Let $\Omega = \mathbf{Diag}(\mathbf{Pr}(M))$ such that $\Omega^{-1} = \mathbf{Diag}(1/\Pr(M = m_1), \ldots, 1/\Pr(M = m_K))$. Then $\mathbf{Pr}(\mathbf{G}|\mathbf{M}) = \Omega^{-1}\mathbf{Pr}(\mathbf{M}, \mathbf{G})$ so that $\mathbf{Pr}(\mathbf{G}|\mathbf{M})^{-1} = \mathbf{Pr}(\mathbf{M}, \mathbf{G})^{-1}\Omega$ can be calculated. In terms of this quantity, the relationship between $\mathbf{p}$ and $\boldsymbol{\pi}$ is thus

$$\boldsymbol{\pi} = \mathbf{Pr}(Y = 1|\mathbf{M}) = \mathbf{Pr}(\mathbf{G}|\mathbf{M}) \times \mathbf{p},$$

so that

$$\mathbf{p} = [\mathbf{Pr}(\mathbf{G}|\mathbf{M})]^{-1} \times \boldsymbol{\pi}.$$

This gives $\mathbf{p}$ as a function of $\mathbf{r}$, θ, and $\boldsymbol{\pi}$. If the marker map is known [and hence $\mathbf{Pr}(\mathbf{M})$ is known], $\mathbf{p}$ is a function of only $\mathbf{r}$ and $\boldsymbol{\pi}$.

Estimation: To estimate recombination ($\mathbf{r}$) and penetrance ($\mathbf{p}$) parameters the invariance property of MLEs can be invoked (Casella and Berger 1990). Thus

$$\hat{\mathbf{p}} = [\mathbf{Pr}(\mathbf{G}|\mathbf{M})]^{-1}\big|_{\mathbf{r}=\hat{\mathbf{r}}} \times \hat{\boldsymbol{\pi}}. \quad (3)$$

The resulting system of equations is linear in $\mathbf{p}$ and $\boldsymbol{\pi}$ and nonlinear in $\mathbf{r}$. Since there are K equations and $K + k$ unknowns, the system is underdetermined. For any fixed $\mathbf{r}$, however, there is a unique and easily obtained solution for $\mathbf{p}$, and, furthermore, the values are subject to constraints, namely $0 \le r_i \le 0.5$ and $0 \le p_{g_j} \le 1$. We use a grid search to step through the interval of possible $\mathbf{r}$ values to obtain sets of solutions for $\mathbf{p}$. In some cases, solutions do not satisfy the biological constraint $0 \le \hat{p}_{g_j} \le 1$ for all g_j and can be discarded, if a filter on the resulting estimates is desired. In other situations, the correct values of some penetrances may be known or separately estimable from previous experimental generations. In these cases estimates can be further constrained. For example, if parental penetrances are known, the values of $\hat{\mathbf{r}}$ and $\hat{\mathbf{p}}$ that minimize the distance between the known values of the parental penetrances and their estimates can be chosen. Other possible solutions to this problem exist; for example, nonlinear programming methods may be applied. We initially explored applying nonlinear programming (NLP) techniques; however, estimation for k loci was not easily implemented.

Premodeling strategy: Generally, mapping data consist of a set of markers and a trait evaluation for each individual in the experimental population. The number of BTL that can be fit to the data depends on the sample size (*i.e.*, the degrees of freedom). If the set of markers is relatively large, as is generally the case, enumerating all possible models becomes impossible, and we need some methodology for reducing the model space. We propose to limit the model space explored by choosing one marker per linkage group. Another strategy for reducing the model space is to limit multiple-marker models to only markers that are significantly associated with the trait on the basis of single-locus models (Kao *et al.* 1999; Carlborg and Andersson 2002). However, this strategy might miss markers without strong main effects that are otherwise involved in epistasis. Another strategy is to examine all possible pairs of loci (Holland *et al.* 2002; Yi and Xu 2002) to reduce the chance of missing a locus with primarily epistatic effect. While it is possible to look at all possible pairs of markers, examining all possible triplets and quadruplets quickly becomes impractical. Additionally, significant model selection bias and uncertainty is introduced (Burnham and Anderson 2002; Bogdan *et al.* 2004).

To avoid stepwise procedures and selection methods based upon pairwise relationships it has been proposed that the relationships among predictor variables can be exploited to reduce model space (Harrell 2001). Fortunately, markers have an inherent relationship among themselves based on genetic distance and form groups of correlated covariates known as linkage groups. By selecting the best marker (or interval) from each linkage group, the dimensionality of the problem is greatly reduced. The criteria for choosing the best among the markers for a linkage group are also possible to explore.

TABLE 2

Simulation conditions: simulation 1

r_1	Genetic model	p_1	p_2
		One-locus simulations	
0.20, 0.30, 0.50	Additive	0.00, 0.20, 0.50	0.80, 1.00

r_1	r_2	Penetrance group	Genetic model	p_{11}	p_{12}	p_{21}	p_{22}
			Two-locus simulations				
0.10, 0.20, 0.30	0.10, 0.20, 0.30	1	Additive	0.20	0.47	0.73	1.00
0.40, 0.50	0.40, 0.50	2	Rec. epistasis 1	0.20	0.20	0.50	1.00
			Rec. epistasis 2	0.20	0.50	0.20	1.00
			Rec. epistasis 3	0.20	1.00	0.20	0.20
			Rec. epistasis 4	0.20	0.20	1.00	0.20
		3	Epistasis 1	0.20	0.50	1.00	0.20
			Epistasis 2	0.50	0.20	0.20	1.00
0.3	0.3		Null model	0.70	0.70	0.70	0.70

r_1	r_2	r_3	Penetrance group	Genetic model	p_{111}	p_{112}	p_{121}	p_{122}	p_{211}	p_{212}	p_{221}	p_{222}
				Three-locus simulations								
0.20	0.20	0.20	1	Additive	0.20	0.31	0.43	0.54	0.66	0.77	0.89	1.00
			2	Rec. epistasis 1	0.20	0.20	0.43	0.54	0.66	0.77	0.89	1.00
			3	Epistasis 1	0.20	0.31	0.43	0.54	0.66	0.77	0.89	0.20

One-locus simulation conditions: one marker (M_1) and one BTL (G_1), where r_1 is the recombination between M_1 and G_1, p_1 is the penetrance of genotype G_1G_1 and p_2 is the penetrance of genotype G_1G_2. Two-locus and three-locus simulation conditions: two markers (M_1M_2) and two BTL (G_1G_2) and three markers ($M_1M_2M_3$) and three BTL ($G_1G_2G_3$), respectively, where r_i is the recombination between M_i and G_i, θ_i is the recombination between M_i and M_{i+1}, three values for θ_i were simulated for each combination of parameters (low to no linkage, medium linkage, and high linkage), and p_{g_j} is the penetrance of genotype g_j.

For simplicity, we choose the marker with the lowest *P*-value. However, if there are unequal amounts of missing data, it would be possible to include the amount of missing data as a criterion for selection. AIC and BIC could also be used in this context. By choosing a marker in each linkage group without regard to the "significance," epistatic loci that show little or no main effect can be detected. The reduction in overall dimensionality reduces the number of models. Thus, the genetic model space can be explored without the assistance of complex searching algorithms and the overall model bias and uncertainty are reduced.

Model selection: To select a model or set of models from among a number of models, standard model selection criteria, AIC (Akaike 1973) and BIC (Schwarz 1978), are often employed. Mallow's C_p (Mallow 1973) is another commonly used criterion that tends to select the the same models as AIC (Quinn and Keough 2002). We explored the behavior of C_p in this context and found it to be very similar to AIC; therefore, we did not include Mallow's C_p in our formal evaluation of model selection criteria.

AIC is a very general methodology based on the theory of optimization where the goal is to select the best approximating model or set of approximating models supported by the empirical data. Furthermore, a small sample AIC (Sugiura 1978), denoted AIC_c, is available to be used when the ratio of sample size (n) to number of parameters (p) is small (*i.e.*, <40) (Burnham and Anderson 2002). In contrast, dimension-consistent criteria (*e.g.*, BIC) assume that one of the models is the true model and is not based in the theory of optimization. Implicit in the assumption that one of the models is the "true" model is that the "truth" is of fairly low dimension (Burnham and Anderson 2002). Asymptotically the BIC will select the true model with probability 1, if that model is in the set. The goal of these criteria (AIC or AIC_c and BIC) is to allow for ranking and comparison of models to separate models that are equally useful from those that are clearly not useful (Burnham and Anderson 2002).

Whichever criterion is applied, models or sets of models need to be delineated for evaluation. When large numbers of models are evaluated, model uncertainty and parameter estimation bias are likely outcomes (Burnham and Anderson 2002). By *selecting one marker per linkage group*, regardless of whether that model is significant, the full set of hierarchical regression models can be fit. This eliminates the need for stepwise procedures. For example, if there are 10 linkage groups and the marker with the strongest individual effect is chosen from each group, then the set of all possible models

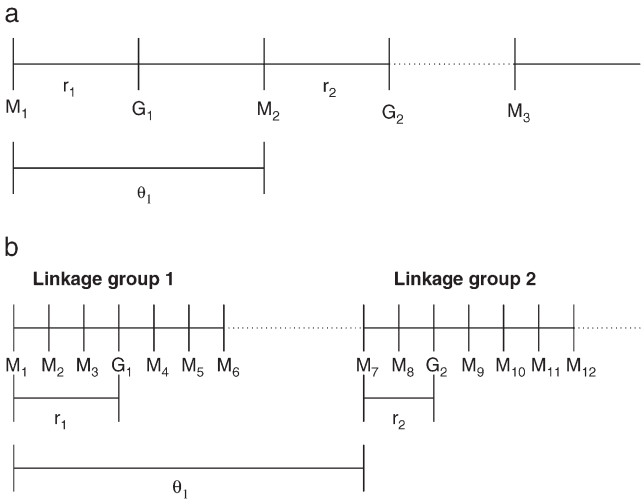

FIGURE 1.—Genetic map of markers for two-locus simulations where markers are denoted M_i and BTL are denoted G_i. (a) Simulation 1: r_1 is the recombination rate between M_1 and G_1, r_2 is the recombination rate between M_2 and G_2, θ_1 is the recombination rate between M_1 and M_2, and M_3 is an unlinked marker. (b) Simulation 2: r_1 is the recombination rate between M_1 and G_1 on linkage group 1, r_2 is the recombination rate between M_7 and G_2 on linkage group 2, and θ_1 is the recombination rate between M_1 and M_7.

includes the 10 single-locus regression models, all 45 two-locus models, all 120 three-locus models, and so on. Thus the limitation in fitting higher-level models is not the ability to search the model space, but rather sample size. With a limited sample size, the addition of marker loci can cause a separation of points, as not enough individuals are observed for all the marker class combinations. For example, to fit four loci in a backcross, we have 16 marker class combinations. If a sample is insufficient to estimate a higher-level model space, models will fail to converge. Although estimates can sometimes be obtained for models that are too large for the data, examination of criteria like BIC will indicate that these models do not fit better than models of lower dimension. If models for up to three loci consistently converge but four-locus models do not, then a total of 175 models will be fit. For each of these 175 models, the model selection criteria are applied and the best models from the entire set are selected. Once a model or set of models has been identified, model tests and parameter estimates can be evaluated.

Simulations: We performed two sets of simulations that we refer to as simulation 1 and simulation 2 (Table 2). In simulation 1 the number of markers was limited, while the number of different genetic models varied widely. In simulation 2 we chose a subset of representative models and then examined the impact of adding a "genome scan."

In simulation 1, we simulated a sample size of 1000 individuals from a backcross population, with 1, 2, and 3 BTL, using $t + 1$ markers. The t markers were linked and adjacent to a BTL with one marker not linked to any BTL (see Figure 1). Since the objective is to study the impact of the genetic model, a large sample size was chosen. When only one BTL locus is truly present ($k = 1$), there is no epistasis.

When more than one locus is involved, the set of genetic models explored in a simulation study is essentially infinite. For convenience we categorized the genetic model space into three groups(see Figure 2):

1. Group 1: additive models, p_{g_j} parameters are all different (equally spaced).
2. Group 2: dominant or recessive epistasis, $p_{12} = p_{22}$ or $p_{21} = p_{22}$ (dominance) or $p_{12} = p_{11}$ or $p_{21} = p_{11}$ (recessive).
3. Group 3: epistasis, $p_{11} - p_{12} \neq p_{21} - p_{22}$.

Loci can have either a weak effect (*e.g.*, $p_{22} - p_{11} = 0.4$) or a strong effect (*e.g.*, $p_{22} - p_{11} = 0.8$) (COHEN 1988).

For two BTL, we explored penetrance models in these three groups of genetic models. On the basis of these results, we selected a model from each of the three groups (additive, a recessive epistasis, or epistasis) for simulations of three BTL (see Figure 3 for three-loci models). We simulated a total of 571 combinations of r and p (see Table 2). For each combination of parameters, 1000 simulation replicates were performed.

We calculated the likelihood-ratio test (LRT) of the correct model compared to the null model to estimate the power to detect BTL for each replicate simulation. The null hypothesis was rejected when the empirical *P*-value for the replicate was less than a nominal significance level of 0.05. Empirical *P*-values were obtained via permutation. The power for the correct model was estimated as the number of times the empirical *P*-value for that replicate was <0.05 divided by the number of replicates.

For each set of simulation conditions, we estimated recombination ($\mathbf{r}$) and penetrance ($\mathbf{p}$) according to the correct model. We used a grid search from $r_i = 0.0$ to $r_i = 0.5$ with a step size of 0.05 for $i = 1$–k. For each replicate and each combination of r's, $\hat{\mathbf{p}}$ was calculated using estimates of recombination $\hat{\theta}_i$ from the data. A set of unconstrained $\hat{\mathbf{r}}$ and $\hat{\mathbf{p}}$ estimates and constrained $\hat{\mathbf{r}}$ and $\hat{\mathbf{p}}$ estimates was generated. Combinations of $\hat{\mathbf{r}}$ and $\hat{\mathbf{p}}$ that did not satisfy $-0.10 \leq \hat{p}_j \leq 1.1$ were discarded. Estimates ($\mathbf{r}$ and $\mathbf{p}$) were averaged to determine an unconstrained estimate for each replicate. Constrained estimates were obtained by selecting the values of $\hat{\mathbf{r}}$ and $\hat{\mathbf{p}}$ that minimized the distance between the simulated values for the parental penetrances and their estimates. If more than one set of $\hat{\mathbf{r}}$ and $\hat{\mathbf{p}}$ had the same distance, the sets were averaged for that replicate.

Following the assessment of power and the evaluation of the estimation procedures, models with differing numbers of BTL loci were fit, with 0, 1, . . . , t, $t + 1$ loci for a total of 2^{t+1} models for each replicate. The cell means model (equivalent to the full factor effects

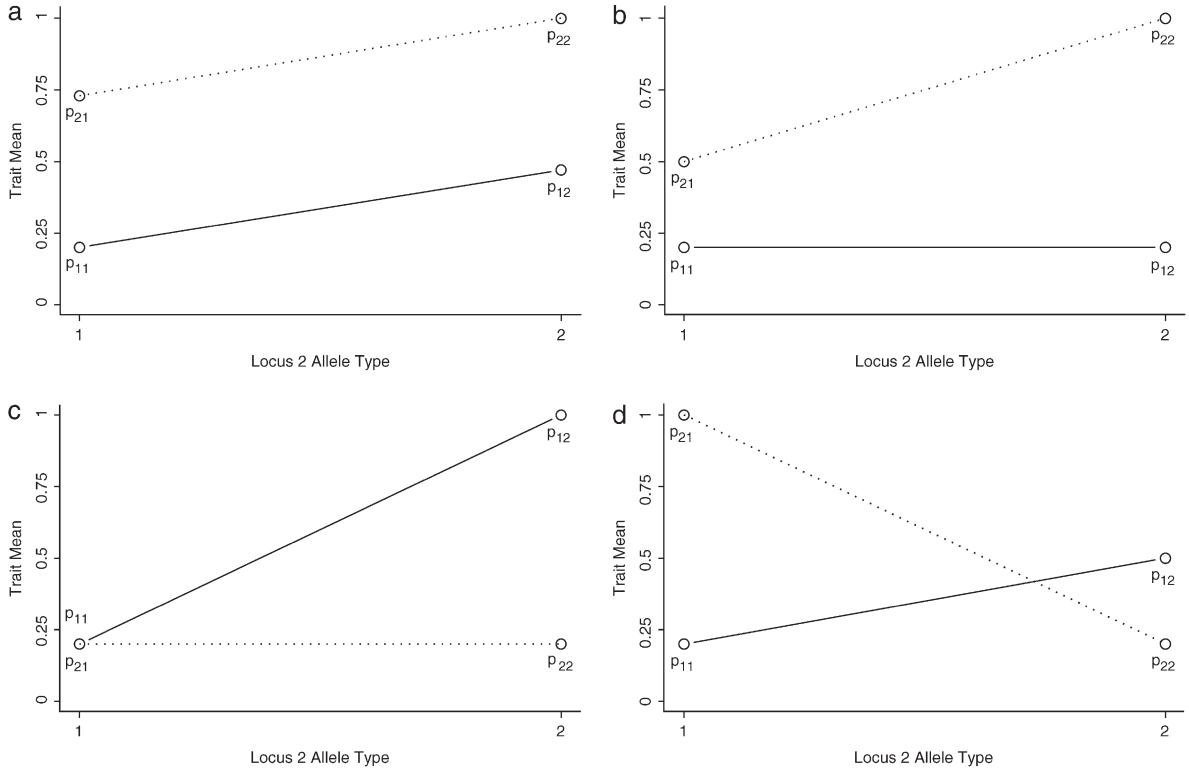

FIGURE 2.—Plots of a penetrance model from each of three simulated groups for two loci. (a) Group 1, additive; (b) group 2, recessive (rec.) epistasis 1; (c) group 2, rec. epistasis 3; (d) group 3, epistasis 1. Solid line, locus 1, allele type 1; dashed line, locus 1, allele type 2, with p_{g_j} representing the penetrance of genotype g_j.

model including interaction terms) was fit. The model selection criteria AIC and BIC were calculated for each of the 2^{t+1} models in a particular replicate. The model with the lowest value for each of the two criteria was determined and counted as a success for that replicate. The proportion of successes for each of the 2^{t+1} models was determined by summing the number of successes for each model divided by the number of replicates.

In simulation 2, we selected four genetic models that were representative of groups 1–3 explored in simulation 1 (see Table 3). For each of these cases, a backcross population with 1000 individuals and 10 linkage groups with 5–20 markers per linkage group for a total of 100 markers was simulated. Two BTL were considered, and the locations of the BTL were determined randomly with the constraint that the two BTL occur on separate linkage groups. In this case, we applied the premodeling strategy of selecting one marker from each linkage group and explored the model selection problem in this context. A single-marker analysis was conducted, and the marker with the lowest *P*-value on each linkage group was selected for further examination without regard to significance. The resulting set of $m = 10$ markers was then used to fit the null model, all 10 single-marker models, all 45 2-marker models, all 120 3-marker models, and all 210 4-marker models (for a total of 386 possible models). The model selection criteria AIC and BIC were calculated for each of the models and the model with the lowest value for each of the two criteria was determined and counted as the selected model for that replicate. The proportion of times the model was selected was determined by summing the number of selections for each model divided by the number of replicates.

RESULTS

Simulation 1: Overall, the proposed maximum-likelihood approach performed well for estimating parameters. As expected, the constrained estimates are closer to the simulated values than the unconstrained estimates. For example, with constrained estimates, for the simulation with $r_1 = 0.10$ and $r_2 = 0.10$ and genetic model recessive (rec.) epistasis 1 from group 2 (see Table 2), estimates were $\hat{r}_1 = 0.12$, $\hat{r}_1 = 0.10$, $\hat{p}_{11} = 0.20$, $\hat{p}_{12} = 0.21$, $\hat{p}_{21} = 0.48$, and $\hat{p}_{22} = 1.00$. The unconstrained estimates for this same simulation were $\hat{r}_1 = 0.13$, $\hat{r}_1 = 0.10$, $\hat{p}_{11} = 0.20$, $\hat{p}_{12} = 0.20$, $\hat{p}_{21} = 0.46$, and $\hat{p}_{22} = 1.01$. For the simulation with $r_1 = 0.10$ and $r_2 = 0.10$ and genetic model epistasis 1 from group 3 (see Table 2), constrained estimates were $\hat{r}_1 = 0.06$, $\hat{r}_1 = 0.11$, $\hat{p}_{11} = 0.20$, $\hat{p}_{12} = 0.47$, $\hat{p}_{21} = 0.89$, and $\hat{p}_{22} = 21$. The unconstrained estimates for this same simulation were $\hat{r}_1 = 0.22$, $\hat{r}_1 = 0.13$, $\hat{p}_{11} = 0.20$, $\hat{p}_{12} = 0.40$, $\hat{p}_{21} = 0.59$, and

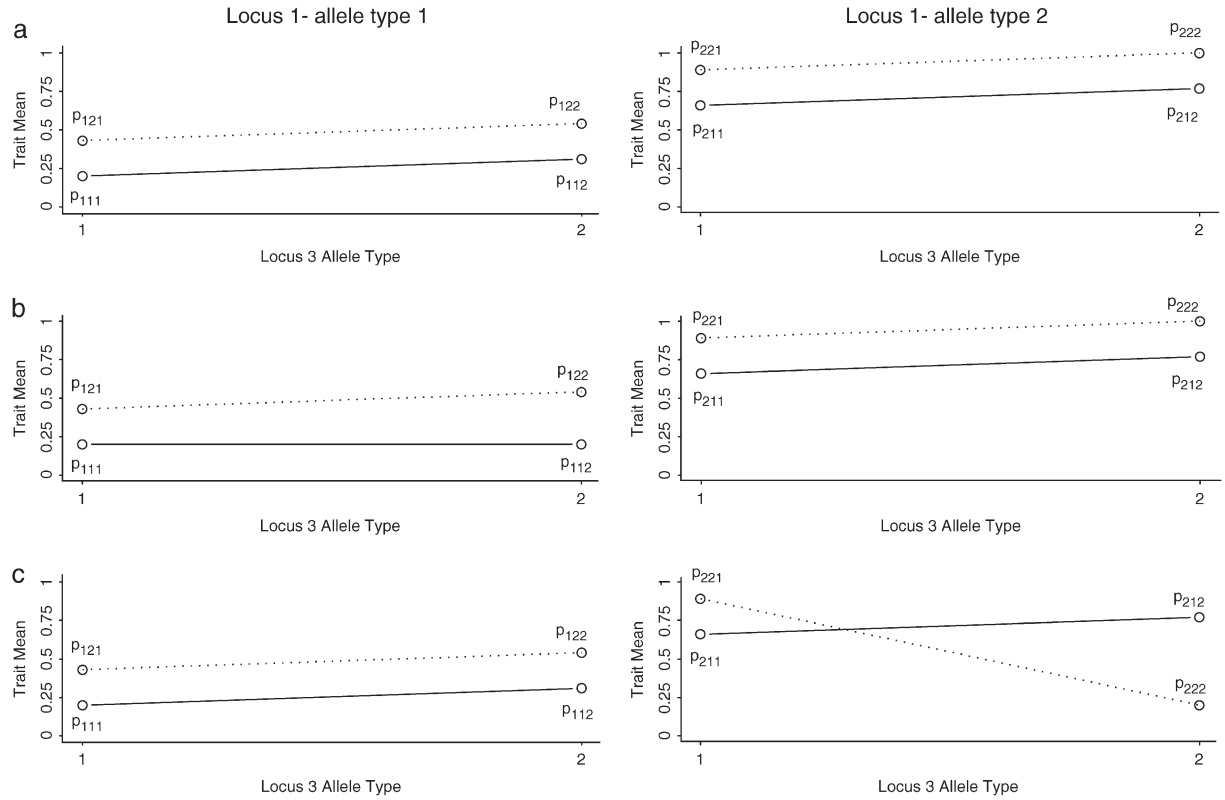

FIGURE 3.—Plots of a penetrance model from each of three simulated groups for three loci. (a) Group 1, additive; (b) group 2, rec. epistasis 1; (c) group 3, epistasis 1. Solid line, locus 2, allele type 1; dashed line, locus 2, allele type 2, with p_{g_j} representing the penetrance of genotype g_j.

$\hat{p}_{22} = 0.12$. Using the median rather than the average of the set of estimates does not improve estimation (results not shown). For each iteration, more than one solution that satisfies the system of equations may be obtained. By definition, all solutions are equally likely. We calculated the average of all equally likely solutions. When estimates are unconstrained, this average will include values that are not biologically meaningful, and when constrained this average will include all values that satisfy biological constraints.

Power for detection of BTL is fairly high for most models examined (see Figure 4). However, the lowest estimate of power observed was 0.43 for the case $r_1 = 0.40$ and $r_2 = 0.40$, for the genetic model epistasis 1 in group 3. As a check of the simulations, we examined the null case, when all penetrance parameters are equal, and achieved the expected nominal significance level as the estimate of power. In BTL mapping, as in QTL mapping, when linkage between the marker and BTL decreases, power decreases. Power for all genetic models, including most epistatic models, is comparable to power for the additive genetic model except when the marker is fairly distant from the BTL locus (r_1 or $r_2 \geq 0.30$). Consistent with the QTL literature we find that power is also dependent on the distance between the marker and the BTL loci and on sample size (results not shown).

Following the exploration of estimation and power,

TABLE 3

Simulation conditions: simulation 2 (two-locus simulations)

r_1	r_2	Penetrance group	Genetic model	p_{11}	p_{12}	p_{21}	p_{22}
0.20	0.20	1	Additive	0.20	0.47	0.73	1.00
		2	Rec. epistasis 1	0.20	0.20	0.50	1.00
		3	Epistasis 1	0.20	0.50	1.00	0.20

Ten linkage groups are shown with 5–20 markers per linkage group equally spaced. r_i is the recombination between M_i and G_i, θ_i is the recombination between M_i and M_{i+1}, and p_j is the penetrance of the jth genotype. r_i values of 0.50 were not simulated for these models.

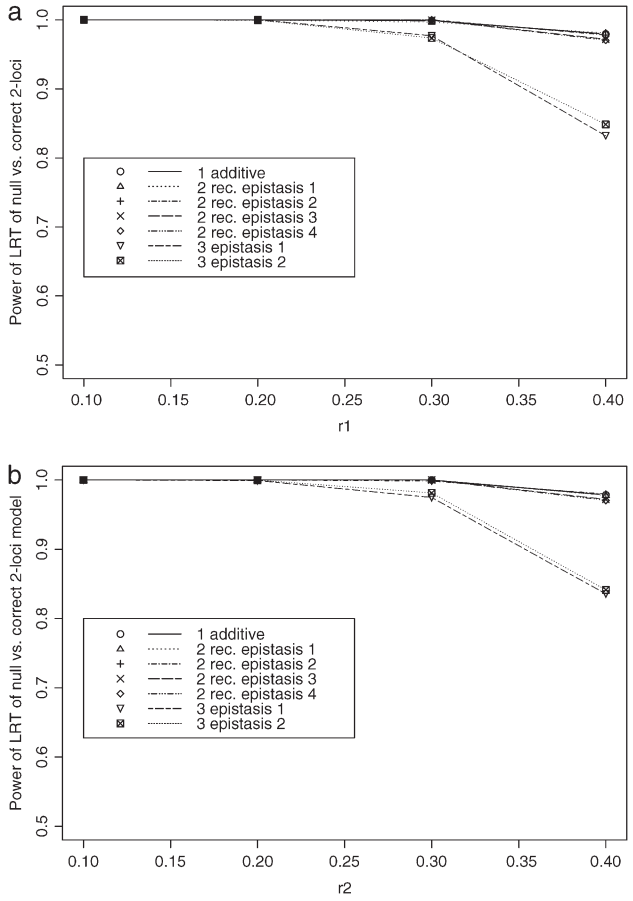

FIGURE 4.—Power for correct two-locus model for each of the simulated penetrance group genetic models when markers M_1 and M_2 are *unlinked* to each other. (a) Power with recombination between M_1 and G_1 (r_1) on the *x*-axis is averaged over recombination between M_2 and G_2. (b) Power with recombination between r_2 on the *x*-axis is averaged over r_1. Penetrance group (1–3) and genetic model are given in the inset.

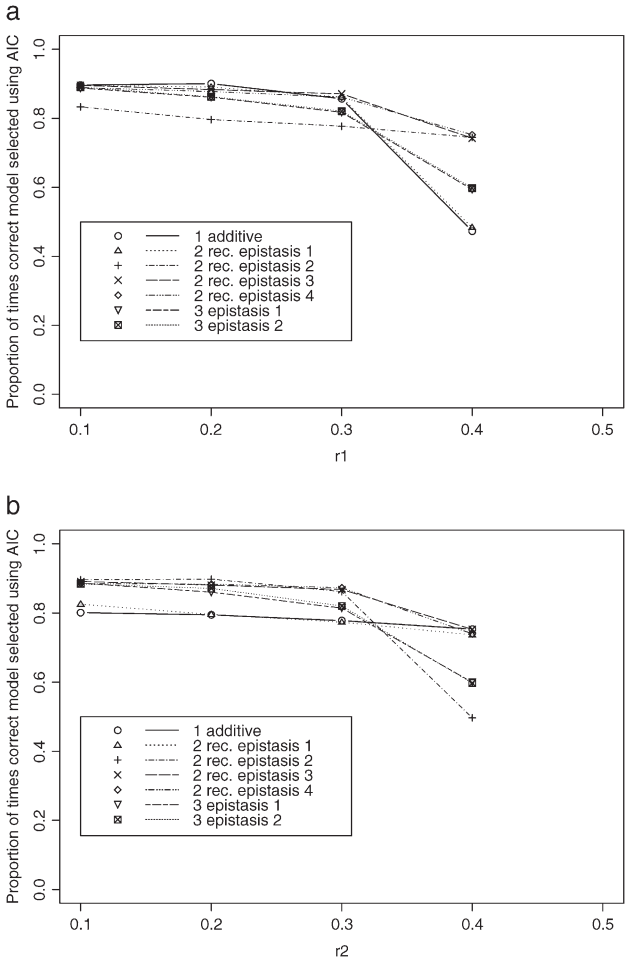

FIGURE 5.—Model selection: proportion of times the correct two-locus model was selected using the AIC when markers M_1 and M_2 are unlinked to each other. (a) Recombination between M_1 and G_1 (r_1) on the *x*-axis is averaged over recombination between M_2 and G_2. (b) Recombination between M_1 and G_1 (r_2) on the *x*-axis is averaged over r_1. Penetrance group (1–3) and genetic model are given in the inset.

the performance of the standard model selection criteria, AIC and BIC, was examined over a wide range of genetic models. As a check, we examined the null cases, where all penetrance parameters are equal or all BTL are unlinked to the markers at hand and, as expected, the null model was typically selected by both criteria. For additive models, selection of the correct model was affected by recombination and the difference between the penetrance parameters. BIC appeared to be more sensitive than AIC to the recombination rate. For example, when one BTL was considered at $r_1 = 0.20$, and the difference between penetrance parameters is large (*i.e.*, 0.80), AIC selects the correct model in 86% of simulations and the BIC selects the correct model in 99.5% of simulations. However, when the recombination rate increases to $r_1 = 0.30$ with the same effect size, AIC selects the correct model in 52% of simulations and the BIC selects the correct model in only 15% of simulations. Epistatic models showed the same trend: as recombination between the marker and the BTL increased, or the difference between the penetrance parameters decreased, the likelihood of choosing the correct model decreased (see Figure 5).

Over all genetic models, AIC tends to select the correct model at a higher rate than BIC for $k = 1$, 2, and 3 simulated BTL whether the markers were linked to each other ($\theta_i < 0.50$) or not ($\theta_i = 0.50$) (see Tables 4 and 5). For two BTL, when the markers were linked ($\theta_i < 0.50$) the correct model was selected 80% of the time in 50% of the simulated scenarios for AIC and 25% for BIC. For unlinked markers ($\theta_i = 0.50$) both AIC and BIC selected the correct model 80% of the time at a higher rate, 73% for AIC and 48% for BIC.

For two-BTL simulations, BIC is more sensitive than AIC to recombination. For example, in the additive model with $r_1 = 0.30$, $\theta_1 = 0.42$, $r_2 = 0.20$ BIC selects the correct model only 39% of the time. When recombination decreases to $r_1 = 0.20$ and all other parameters remain the same, BIC selects the correct model 71% of

TABLE 4

Simulation 1: the proportion of times that the model was selected for specified criteria (AIC, BIC) for the null model (no loci) and all one-, two-, and three-locus models, where the correct model is in italics

		AIC								BIC							
			One locus			Two loci			Three loci:		One locus			Two loci			Three loci:
Penetrance group	Genetic model	Null	M_1	M_2	M_3	M_1M_2	M_1M_3	M_2M_3	$M_1M_2M_3$	Null	M_1	M_2	M_3	M_1M_2	M_1M_3	M_2M_3	$M_1M_2M_3$
							$r_1 = 0.30$, $\theta_1 = 0.42$, $r_2 = 0.20$										
1	Additive	0.00	0.01	0.00	0.00	*0.87*	0.00	0.00	0.12	0.00	0.29	0.00	0.00	*0.71*	0.00	0.00	0.00
2	Rec. epistasis 1	0.00	0.01	0.00	0.00	*0.92*	0.00	0.00	0.08	0.00	0.23	0.00	0.00	*0.77*	0.00	0.00	0.00
	Rec. epistatis 2	0.00	0.00	0.00	0.00	*0.87*	0.00	0.00	0.13	0.00	0.00	0.00	0.00	*1.00*	0.00	0.00	0.00
	Rec. epistasis 3	0.00	0.00	0.02	0.00	*0.88*	0.00	0.01	0.09	0.00	0.00	0.38	0.00	*0.62*	0.00	0.00	0.00
	Rec. epistasis 4	0.00	0.00	0.02	0.00	*0.89*	0.00	0.00	0.09	0.01	0.00	0.35	0.00	*0.65*	0.00	0.00	0.00
3	Episatsis 1	0.00	0.00	0.00	0.00	*0.90*	0.00	0.00	0.09	0.08	0.00	0.10	0.00	*0.82*	0.00	0.00	0.00
	Epistasis 2	0.00	0.00	0.00	0.00	*0.88*	0.00	0.00	0.12	0.00	0.02	0.00	0.00	*0.98*	0.00	0.00	0.00
							$r_1 = 0.30$, $\theta_1 = 0.42$, $r_2 = 0.20$										
1	Additive	0.00	0.07	0.00	0.00	*0.83*	0.01	0.00	0.09	0.00	0.61	0.00	0.00	*0.39*	0.00	0.00	0.00
2	Rec. epistasis 1	0.00	0.05	0.00	0.00	*0.82*	0.02	0.00	0.11	0.00	0.55	0.00	0.00	*0.45*	0.00	0.00	0.00
	Rec. epistasis 2	0.00	0.00	0.00	0.00	*0.89*	0.00	0.00	0.11	0.00	0.00	0.00	0.00	*1.00*	0.00	0.00	0.00
	Rec. epistasis 3	0.00	0.00	0.01	0.00	*0.91*	0.00	0.01	0.08	0.03	0.02	0.22	0.00	*0.73*	0.00	0.00	0.00
	Rec. epistasis 4	0.00	0.00	0.01	0.00	*0.88*	0.00	0.00	0.11	0.03	0.03	0.22	0.00	*0.73*	0.00	0.00	0.00
3	Epistasis 1	0.00	0.00	0.01	0.00	*0.87*	0.00	0.00	0.11	0.23	0.03	0.13	0.00	*0.61*	0.00	0.002	0.00
	Epistasis 2	0.00	0.00	0.00	0.00	*0.91*	0.00	0.00	0.09	0.00	0.17	0.00	0.00	*0.82*	0.00	0.00	0.00

One thousand individuals, 1000 replicates, and two loci were used. M_i denotes the marker at locus i and G_i denotes BTL i, r_i is the recombination between M_i and G_i, and θ_1 is the recombination between M_1 and M_2 (see simulated map in Figure 1).

TABLE 5

Simulation 1: proportion of times the model was selected for specified criteria (AIC, BIC) for the null model (no loci) and the one-, two-, three-, and four-locus models

Penetrance group	Genetic model	AIC					BIC				
		Null	One locus	Two loci	*Three loci*	Four loci	Null	One locus	Two loci	*Three loci*	Four loci
		$r_1 = 0.20, \theta_1 = 0.20, r_2 = 0.20, \theta_2 = 0.20, r_3 = 0.20$									
1	Additive	0.00	0.00	0.57	*0.31*	0.10	0.03	0.00	0.02	0.97	*0.00*	0.00	0.00
2	Rec. epistasis 1	0.00	0.00	0.68	*0.21*	0.09	0.02	0.00	0.03	0.97	*0.00*	0.00	0.00
3	Epistasis 1	0.00	0.00	0.06	*0.89*	0.01	0.04	0.00	0.02	0.93	*0.07*	0.00	0.00
		$r_1 = 0.20, \theta_1 = 0.50, r_2 = 0.20, \theta_2 = 0.50, r_3 = 0.20$									
1	Additive	0.00	0.001	0.44	*0.47*	0.06	0.03	0.00	0.15	0.85	*0.00*	0.00	0.00
2	Rec. epistasis 1	0.00	0.00	0.57	*0.34*	0.08	0.02	0.00	0.04	0.96	*0.00*	0.00	0.00
3	Epistasis 1	0.01	0.01	0.11	*0.83*	0.02	0.02	0.00	0.65	0.29	*0.06*	0.00	0.00

For incorrect models, proportions are the proportion of all possible *t*-locus models. The correct model is in italics and the proportion of other three-locus models selected is in regular type. One thousand individuals, 1000 replicates, and three loci were used. M_i denotes the marker at locus i and G_i denotes BTL locus i, r_i is the recombination between M_i and G_i, and θ_i is the recombination between M_i and M_{i+1} (see simulated map in Figure 1).

the time (see Table 4). This is intuitively logical, since the distance between the BTL and the marker increases, and the effect of the penalty for the BIC is more severe than the effect of the penalty for the AIC, making the BIC more sensitive than AIC to recombination distance.

Simulation 2: For simulation 2, the focus was on a subset of genetic models, one from each of the penetrance groups where a large number of extra markers were included. This simulation provides an opportunity to examine the performance of the AIC and BIC in a more realistic data analytic setting. In these simulations, the BIC far outperformed the AIC (see Table 6) and resulted in a higher likelihood of choosing the correct model. AIC tended to select models of too high dimension. The behavior of the AIC was dramatically different between simulations 1 and 2 (see Tables 4 and 6). The BIC performed similarly in genetic models from group 1 and group 2 while the performance of the BIC in a genetic model from group 3 was affected by the additional marker. This change is not nearly as dramatic as the change for the AIC. However, for the BIC, the genetic model affected whether the two-BTL model that included the simulated BTL was selected or not. For example, for the genetic model epistasis 1 from group

TABLE 6

Simulation 2: proportion of times the model was selected for specified criteria (AIC, BIC) for the null model (no loci) and the one-, two-, three-, and four-locus models with 10 linkage groups with 5–20 markers per linkage group

Penetrance group	Genetic model	AIC						BIC					
		Null	One locus	*Two loci*		Three loci	Four loci	Null	One locus	*Two loci*		Three loci	Four loci
		$r_1 = 0.20, \theta_1 = 0.32, r_2 = 0.20, r_M = 0.30$											
1	Additive	0.00	0.00	*0.07*	0.00	0.40	0.52	0.00	0.26	*0.71*	0.03	0.00	0.00
2	Rec. epistasis 1	0.00	0.00	*0.07*	0.01	0.40	0.51	0.00	0.22	*0.77*	0.01	0.00	0.00
3	Epistasis 1	0.00	0.00	*0.01*	0.03	0.75	0.21	0.05	0.41	*0.07*	0.47	0.00	0.00
		$r_1 = 0.20, \theta_1 = 0.50, r_2 = 0.20, r_M = 0.30$											
1	Additive	0.00	0.00	*0.09*	0.00	0.46	0.45	0.00	0.03	*0.96*	0.01	0.00	0.00
2	Rec. epistasis 1	0.00	0.00	*0.06*	0.00	0.52	0.42	0.00	0.03	*0.97*	0.00	0.00	0.00
3	Epistasis 1	0.00	0.00	*0.09*	0.00	0.48	0.43	0.00	0.01	*0.97*	0.01	0.00	0.00

The correct model is in italics and the proportion of other two-locus models selected is in regular type. One thousand individuals, 500 replicates, and two loci were used. M_i denotes the marker at locus i and G_i denotes BTL i, r_i is the recombination between M_i and G_i, θ_1 is the recombination between M_1 and M_2, and r_M is recombination between markers on linkage groups (in these simulations, equally spaced). See simulated map in Figure 1.

TABLE 7

Multiple-locus models with the lowest AIC_c criterion for each BTL marker name from the linkage group given with linkage group number in parentheses

Linkage group								
BTL 1	BTL 2	BTL 3	AIC_c	δ_{AIC_c}	AIC	δ_{AIC}	BIC	δ_{BIC_c}
accaag15 (1)	accaag8 (12)	acgaca20 (17)	32.40	0.00	23.83	0.00	36.73	0.00
agcagc11 (6)	accaag8 (12)	acgaca20 (17)	32.53	0.13	23.96	0.13	36.87	0.13
acgact13 (9)	accaag8 (12)	acgaca20 (17)	32.53	0.13	23.96	0.13	36.87	0.13
accaag8 (12)	acgaca20 (17)	NA	32.76	0.36	30.36	6.53	37.53	0.79
agcagc5 (3)	accaag8 (12)	acgaca20 (17)	36.50	4.10	27.93	4.10	40.83	4.10
accaag8 (12)	acgaca20 (17)	agcaag1 (30)	36.50	4.10	27.93	4.10	40.83	4.10
accaag8 (12)	acgaca20 (17)	agcccg7 (36)	36.50	4.10	27.93	4.10	40.83	4.10

AIC and BIC criteria are also shown. δ_i is the difference between the model with the lowest criterion value and the criteria values of the model in that row.

3, the BIC selected a two-BTL model in 54% of the cases. However, the correct two-BTL model was selected in only 7% of the cases.

***O. mykiss* data analysis:** Doubled haploids, produced by androgenesis in the second generation from a cross between two clonal lines, were used for a genetic analysis of *C. shasta* resistance. *C. shasta* is a myxozoan parasite that has a two-stage life cycle. One stage is completed in a polychaete worm, *Manyukia speciosa*, and actinospores are released to the water and infect the intestinal tracts of trout, where the organism continues development, producing myxospores that are evident by intestinal scrapings. The complete experiment is described in Nichols *et al.* (2003). Briefly, subyearling doubled haploids were exposed in live cages *in situ* to a pathogen in the Willamette River for 4 days in September 2000. Following the exposure, fish were maintained in flow-through systems at the Center for Disease Research hatchery at Oregon State University. Fish were monitored daily where mortalities were removed, recorded, a fin clip taken, and identification number assigned for genetic analysis. Evidence of *C. shasta* spores was evaluated from intestinal scrapings of each individual. The study was terminated 103 days postexposure and fish still alive were labeled survivors and subsequently euthanized with a lethal dose of anesthetic (MS-222, Argent Laboratories), fin-clipped, assigned individual identification numbers, and evaluated for presence of *C. shasta* spores in the intestine. Only mortalities that died from *C. shasta* infection, as evidenced by the presence of *C. shasta* spores in the intestines, were used for genetic analysis of resistance. None of the surviving fish exhibited *C. shasta* spores from intestinal scrapings. Amplified fragment length polymorphic (AFLP) markers were employed to genotype individuals for construction of a genetic linkage map and genetic analysis of *C. shasta* resistance, as previously described (Nichols *et al.* 2003). Three hundred thirteen markers for 45 segregants were mapped and resulted in 38 linkage groups. The number of AFLP markers in comparison to the sample size is very large.

Three hundred thirteen single-marker models were investigated and the marker with the lowest *P*-value on each of 38 linkage groups was selected for inclusion in the multiple-loci models. Of the 45 segregants, 31 had data for the entire set of 38 selected markers. On the basis of this, we considered the 31 segregants for which complete marker data were available. Using the 38 markers on the 31 segregants, all one-locus and two- and three-loci models were investigated (38 one-locus models, 703 two-loci models, and 8436 three-loci models) using PROC BTL (see appendix for PROC BTL). The AIC_c criteria were used for model selection because of the small sample size. For each of the models the AIC, BIC, and AIC_c were calculated and the best models were selected. The best models were used as input for PROC BTL to estimate the recombination and penetrance parameters.

The 38 single-BTL models, 703 distinct two-BTL models, and 8436 distinct three-BTL models were fit; 702 of 703 two-BTL models converged. The best models based on AIC_c, AIC, and BIC are shown in Table 7. The difference between the lowest model selection criterion value for a particular model and the model in the set with the lowest model selection value is denoted as δ. Of the 8436 models fit, 14% failed to converge, most likely due to the limited degrees of freedom (sample size).

The inclusion of markers *accaag8* and *acgaca20* on linkage groups OC21 and OC27 is statistically accurate according to the model selection results. Six sets of estimated recombination and penetrance parameters from the resulting two-locus model are within the range -0.10–1.1 for the penetrance parameters. Estimates for r_1 and r_2 were very small, ranging from 0.00 to 0.05. Estimates for p_{11} ranged from 0.39 to 0.41, for p_{12} ranged from -0.06 to 0.00, for p_{21} ranged from 0.22 to 0.26,

and for p_{22} ranged from 1.03 to 1.08. The small sample combined with the observation that none of the individuals with allele 1 of marker *accaag8* and allele 2 of marker *acgaca20* survived made the addition of a third locus unwise from an estimation perspective (marker class means $\pi_{11} = 0.097$, $\pi_{12} = 0.00$, $\pi_{21} = 0.065$, and $\pi_{22} = 0.26$).

DISCUSSION

This article presents a general likelihood for multiple BTL. The likelihood formulation presented here is similar to that employed by Yi and Xu (2002) with the exception that their liability function is replaced by our single penetrance parameter. Since the estimation of the liability function is computationally challenging, and the methods employed are often sensitive to the choice of this function, our approach greatly simplifies the likelihood and corresponding evaluation process. By choosing one marker per linkage group in a premodeling step, we greatly reduce the model space and avoid stepwise model selection and complicated searching algorithms. Rather than choosing only markers significant in the single-locus models or examining all possible pairs of loci, the relationships (linkage) between markers can be exploited to choose the best locus for each linkage group. This reduces the model space and the impact of model selection upon the subsequent estimation and testing procedures.

While we focus on selection of a single marker in a linkage group, the idea of reducing the marker set can be applied more broadly. For example, in cases where the linkage group may itself be large, the best marker for some fixed genetic distance may be chosen. Alternatively, two or three markers per linkage group may be selected.

In the first simulation, epistatic models are easier to select correctly than the strictly additive model. Initially this was a surprising result but when a fully additive model is considered, with the restriction of the parameter space for the penetrance parameters, $0 \leq p_j \leq 1$, the marginal effect of any one locus is small. This is what is predicted by Fisher's infinitesimal model with a large number of loci. The extension of this idea will be true in quantitative traits as well if the range of the trait values is restricted. In contrast, epistasis restricts the parameters such that several of the penetrances are equal. The consequence of this is larger marginal effects of individual loci. This underscores the importance of fitting models that include epistatic terms as well as main effects.

The effect of linkage between the BTL changes the performance of the selection criteria. For additive models with recessive epistasis (groups 1 and 2) the influence of linkage among BTL improves model selection. For epistasis that considers two loci in model selection it is more difficult when BTL are linked and for the recessive models in our simulations, performance of the BIC decreases while the AIC remains approximately the same. Examining these models closely, we see that the varying difficulty can be explained by considering the penetrance parameters as mixing parameters and examining the relative effect size difference between the loci.

The goals and results of simulation 2 are markedly different. Unlinked BTL are easier to identify than linked BTL simply because they are considered independently. This effect can be diminished by expanding the model search space once the best model or set of models has been selected from among the restricted set. Examining models that increase the number of loci by adding loci linked to loci already included in the best model will allow for additional opportunities to detect linked BTL, while still restricting the model space to a manageable number of loci.

The comparison between simulation 1 and simulation 2 underscores the main differences among the two criteria examined. The AIC selects the best "approximating" model for the data, and in cases where few markers are available, these are often the correct selections. In the case of a genome scan, this will result in the addition of loci, particularly in the case of linked BTL. The BIC will more often choose the right model among a large set of models when the true model is of relatively low dimension and is included in the set of models to select. In the case of the genome scan the BIC has a larger penalty and thus more often chooses a model of appropriate or lower dimension. However, when the number of loci examined is limited, the penalty for the BIC forces models of too low a dimension to be selected. Bogdan *et al.* (2004) propose a modification to the BIC that accommodates the dimensionality of the BTL application and that could be extended to apply here and perhaps mitigate this finding.

In the analysis of the *O. mykiss* there were a fair number of missing marker data. For the purposes of comparison of the techniques explored in this article, the maximum set of complete data was chosen. This is because one of the main assumptions of both AIC and BIC is a constant sample size. Changing the sample size between models will adversely affect the model selection process and because of the penalty term, especially with respect to the BIC, changing the criterion between models will result in changes in the formulation of the likelihood function. As an additional criterion in the premodeling strategy, one might group markers in the linkage group into a set of best markers and then among those markers choose the marker (or interval) with the most complete data. Furthermore, methods that impute the value of missing marker data show promise to reduce the impact of missing marker data. In addition to the missing marker data, the sample size for these data is exceedingly small. The size is so small that inferences drawn from these data are by necessity suggestive, and further experiments would need to be done to make any defini-

tive conclusions. In general, the small sample correction of the AIC (*e.g.*, AIC_c) is considered preferable compared to invoking the asymptotic behavior of the AIC and BIC. Of note is the fact that the three criteria selected the same set of models with similar ranking among models. Examining the set of models that have similar AIC, AIC_c, and BIC values, it is apparent that linkage groups OC21 and OC27 are a common theme. This is particularly interesting as linkage group OC27 had no significant BTL in the single-marker analysis. This points to the possible identification of an epistatic effect in the absence of a significant main effect for that locus. Additional BTL may be located on linkage groups OC7,OC13,OC15,OC30,OC-a, and OC-b, but the joint estimation of parameters in models of this dimension for this small sample size is not recommended.

The typical treatment of binary traits has restricted the use of these data to single-marker analyses. What we propose here is to acknowledge the full depth of binary traits by allowing a modeling strategy that accommodates the potential for epistasis while being aware of the computational challenges that are present in high-dimensional model spaces. By changing the parameterization of the likelihood function to include marker class means, the estimation of penetrance can be obtained. We implement a grid search technique to obtain the solution. There are other potential solutions to this system of nonlinear equations, but these present a complex numerical problem that is a subject of future work. We have provided an easily accessible procedure in SAS that allows multiple-BTL mapping under a wide range of strategies for the purpose of providing a tool that scientists can use with ease and flexibility.

Even though specific model selection criteria (BIC, AIC, and AIC_c) are employed to evaluate the models that result from the model selection procedure proposed, other criteria could easily be used in conjunction with the premodeling strategy proposed. The optimal criterion for model selection is an open and exciting area of research. In the situation presented here the issue is further complicated by constraining the parameter space, which in turn makes proper evaluation of the correct model choice more difficult than it may otherwise appear.

This work is supported by National Science Foundation grant DBI 98-08026/00-96044 (L.M.M., C.J.C., R.W.D.), National Institutes of Health grants NIA-AG16996 (L.M.M.) and 2G12RR003048 (L.M.M.), U.S. Department of Agriculture (USDA) grant 98-35300-6173 (R.W.D.), USDA-Initiative for Future Agriculture and Food Systems grant N0014-94-1-0318 (R.W.D., L.M.M.), and a Veterans Affairs Health Services Research Postdoctoral Fellowship (C.J.C.).

LITERATURE CITED

Akaike, H., 1973 Information theory as an extension, pp. 267–281 in *Second International Symposium on Information Theory*, edited by B. Petrov and F. Csaki. Akademiai Kiado, Budapest.

Ball, R. D., 2001 Bayesian methods for quantitative trait loci mapping based on model selection: approximate analysis using the Bayesian information criterion. Genetics **159:** 1351–1364.

Bogdan, M., J. K. Ghosh and R. W. Doerge, 2004 Modifying the Schwarz Bayesian information criterion to locate multiple interacting quantitative trait loci. Genetics **167:** 989–999.

Broman, K. W., and T. P. Speed, 2002 A model selection approach for the identification of quantitative trait loci in experimental crosses. J. R. Stat. Soc. Ser. B **64:** 641–656.

Burnham, K. P., and D. R. Anderson, 2002 *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, Ed. 2. Springer, Berlin/Heidelberg, Germany/New York.

Carlborg, O., and L. Andersson, 2002 Use of randomization testing to detect multiple epistatic QTL. Genet. Res. **79:** 175–184.

Carlborg, O., L. Andersson and B. Kinghorn, 2000 The use of a genetic algorithm for simultaneous mapping of multiple interacting quantitative trait loci. Genetics **155:** 2003–2010.

Casella, G., and R. L. Berger, 1990 *Statistical Inference*. Wadsworth & Brooks/Cole, Pacific Grove, CA.

Coffman, C. J., R. W. Doerge, M. L. Wayne and L. M. McIntyre, 2003 Intersection tests for single marker QTL analysis can be more powerful than two marker QTL analysis. BMC Genet. **4:** 10 (http://www.biomedcentral.com/1471-2156/4/10).

Cohen, J., 1988 *Statistical Power Analysis for the Behavioral Sciences*, Ed. 2. Lawrence Earlbaum Associates, Hilldale, NJ.

Doerge, R. W., 2001 Mapping and analysis of quantitative trait loci in experimental populations. Nat. Rev. Genet. **3:** 43–52.

Falconer, D. S., and T. F. C. Mackay, 1996 *Introduction to Quantitative Genetics*, Ed. 4. Longman, Essex, UK.

Gauderman, W. J., and D. C. Thomas, 2001 The role of interacting determinants in the localization of genes, pp. 393–412 in *Advances in Genetics, Vol. 42: Genetic Dissection of Complex Traits*, edited by D. C. Rao and M. A. Province. Academic Press, San Diego.

Haley, C., and S. Knott, 1992 A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity **69:** 315–324.

Harrell, F. E., 2001 *Regression Modeling Strategies With Applications to Linear Models, Logistic Regression, and Survival Analysis*. Springer, New York.

Hartl, D., and E. Jones, 2001 *Genetics: Analysis of Genes and Genomes*. Jones & Bartlett, Sudbury, MA.

Holland, J. B., V. A. Portyanko, D. L. Hoffman and M. Lee, 2002 Genomic regions controlling vernalization and photoperiod responses in oat. Theor. Appl. Genet. **105:** 113–126.

Jannink, J., and R. Jansen, 2001 Mapping epistatic quantitative trait loci with one-dimensional genome searches. Genetics **157:** 445–454.

Jansen, R. C., 1992 A general mixture model for mapping quantitative trait loci by using molecular markers. Theor. Appl. Genet. **85:** 252–260.

Jansen, R. C., 1993 Interval mapping of multiple quantitative trait loci. Genetics **135:** 205–211.

Jansen, R. C., and P. Stam, 1994 High resolution of quantitative traits into multiple loci via interval mapping. Genetics **136:** 1447–1455.

Kao, C.-H., Z-B. Zeng and R. D. Teasdale, 1999 Multiple interval mapping for quantitative trait loci. Genetics **152:** 1203–1216.

Kilpikari, R., and M. J. Sillanpää, 2003 Bayesian analysis of multilocus association in quantitative and qualitative traits. Genet. Epidemiol. **25:** 122–135.

Kutner, M. H., C. J. Nachtsheim, J. Neter and W. Li, 2004 *Applied Linear Statistical Models*, Ed. 5. McGraw-Hill Irwin, New York.

Lander, E. S., and D. Botstein, 1989 Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics **121:** 185–199.

Lynch, M., and B. Walsh, 1998 *Genetics and Analysis of Quantitative Traits*. Sinauer Associates, Sunderland, MA.

Mallow, C. L., 1973 Some comments on C_p. Technometrics **12:** 591–612.

McIntyre, L. M., C. J. Coffman and R. W. Doerge, 2001 Detection and localization of a single binary trait locus in experimental populations. Genet. Res. **78:** 79–92.

Nichols, K., J. Bartholomew and G. H. Thorgaard, 2003 Mapping multiple genetic loci associated with *Ceratomyxa shasta* resistance in *Oncorhynchus mykiss*. Dis. Aquat. Org. **56** (2): 145–154.

Ott, J., 1991 *Analysis of Human Genetic Linkage*. Johns Hopkins University Press, Baltimore.

Quinn, G. P., and M. J. Keough, 2002 *Experimental Design and Data Analysis for Biologists.* Cambridge University Press, Cambridge/London/New York.

Satagopan, J. M., B. S. Yandell, M. A. Newton and T. C. Osborn, 1996 A Bayesian approach to detect quantitative trait loci Markov chain Monte Carlo. Genetics **144:** 805–816.

Schwarz, S. L., 1978 Estimating the dimension of a model. Ann. Stat. **6:** 461–464.

Sen, S., and G. A. Churchill, 2001 A statistical framework for quantitative trait mapping. Genetics **159:** 371–387.

Siegmund, D. O., 2004 Model selection in irregular problems: applications to mapping quantitative trait loci. Biometrika **91:** 785–800.

Sillanpää, M. J., and E. Arjas, 1998 Bayesian mapping of multiple quantitative trait loci from incomplete inbred line cross data. Genetics **148:** 1373–1388.

Sillanpää, M. J., and J. Corander, 2002 Model choice in gene mapping: what and why. Trends Genet. **18:** 301–307.

Simonsen, K. L., 2004 A probability model for the inheritance of binary traits. Technical Report Series tr03–04. Purdue University Statistics Department, West Lafayette, IN.

Sugiura, N., 1978 Further analysis of data by Akaike's information criterion and finite corrections. Commun. Stat. Theor. Methods **7:** 13–26.

Thompson, E. A., 1998 Inferring gene ancestry: estimating gene descent. Int. Stat. Rev. **66:** 29–40.

Whittaker, J. C., R. Thompson and P. M. Visscher, 1996 On the mapping of QTL by regression of phenotypes on marker type. Heredity **77:** 23–32.

Xu, S., 1996 Computation of the full likelihood function for estimating variance at a quantitative trait locus. Genetics **144:** 1951–1960.

Xu, S., 2003 Estimating polygenic effects using markers of the entire genome. Genetics **163:** 789–801.

Xu, S., and W. R. Atchley, 1996 Mapping quantitative trait loci for complex binary diseases using line crosses. Genetics **143:** 1417–1424.

Yi, N., and S. Xu, 2000 Bayesian mapping of quantitative trait loci for complex binary traits. Genetics **155:** 1391–1403.

Yi, N., and S. Xu, 2002 Mapping quantitative trait loci with epistatic effects. Genet. Res. **79:** 185–198.

Yi, N., S. Xu, V. George and D. B. Allison, 2004 Mapping multiple quantitative trait loci for ordinal traits. Behav. Genet. **34:** 3–15.

Zeng, Z-B., 1994 Precision mapping of quantitative trait loci. Genetics **136:** 1457–1468.

Zeng, Z-B., C.-H. Kao and C. J. Basten, 2000 Estimating the genetic architecture of quantitative traits. Genet. Res. **74:** 279–289.

Communicating editor: J. B. Walsh

APPENDIX: SAS PROC BTL

The main components of PROC BTL are the *Marker*, *Model*, and *Parmest* statements. The marker/marker recombination parameters (θ) can be entered directly by the user or can be calculated from a marker map data set with a user-chosen map function (Haldane or Kosambi). Map information is necessary for the implementation of the premodeling strategy described above. However, PROC BTL does not require the map order information.

In the *Marker* statement, all markers that the user wishes to evaluate are listed, and SAS performs the regression of different combinations of marker variables against the trait variable according to criteria specified in the *Model* statement. *Marker* effects are automatically generated by PROC BTL for inclusion in the *Model* equation. Additional variables (covariates) can be added to the model as fixed or random effects. Repeated measures can also be specified. Random effects and repeated measures are specified according to the convention of PROC MIXED. Output tables will be generated by SAS that contain the marker effects along with various model statistics including the likelihood-ratio test of the full model *vs.* the null model, AIC or AIC_c, BIC, and other information criteria. In addition to choosing the best marker in user-defined groups, as proposed in this article, standard stepwise procedures are also available.

The *Parmest* statement fits a BTL model that has a putative BTL to the right of each marker. A specific set of marker/BTL recombination parameters (**r**) can be chosen by the user for each marker in the model, or a grid search can be performed over a range of possible **r** specified by the RSTART and REND options for each marker with default values of $0 \leq r \leq 0.5$. The grid size is specified by the GRID option with a default value of 0.1. If a map is specified, then that map is used for marker/marker recombination parameters (θ). If no map is specified, then marker/marker recombination (θ) is estimated from the data assuming the order of loci specified in the *Marker* statement as the correct order. PROC BTL will calculate the matrix $([\mathbf{Pr}(G|M)]^{-1}|_{\mathbf{r}=\hat{\mathbf{r}}})$ for the set of $\hat{\mathbf{r}}$ in the interval. The $\hat{\mathbf{p}}$ are then obtained by multiplying this matrix by the vector of marker class means (see Equation 3). If the $\hat{\mathbf{p}}$ are in the range specified by the options PMIN and PMAX, or in the default range of $0 \leq \hat{p}_j \leq 1$, then a potential BTL is determined. Confidence limits can be calculated for each $\hat{p}_j$ using the bootstrap method with the option BOOT in the *PARMEST* statement. Complete documentation for PROC BTL is available at http//:www.genomics.purdue.edu/services/software/btl.

```
proc btl data=Marker.input2 map=marker.map outstat=toutput;
     marker m1-m299 m301-m313 /all=1 best=1 mc=p group=chromosome;
   model surv=;
run;

ods html body="MC.htm" frame="MCframe.htm" contents="MCcontents";
   proc print data=output;
   title 'Selected One-Marker Models';
   run;
ods html close;
```

```
proc btl data=Marker.input2 map=marker.map outstat=output2;
   marker m5 m12 m14 m16 m19 m27 m34 m43 m60 m66 m69 m75 m95 m97 m117 m118 m135 m158 m180
      m184 m191 m207 m221 m226 m228 m231 m233 m242 m245 m247 m252 m253 m256 m264 m267
      m268 m294 m310/all=2 mc=AICC;
   model surv=;
run;

ods html body="MC.htm" frame="MCframe.htm" contents="MCcontents";
   proc print data=output2;
   title 'Selected Two-Marker Models';
   run;
ods html close;

proc btl data=marker.input2 map=marker.map output=output3;
   marker m12 m17;
   model surv=;
   parmest /cross=B gen=1 grid=.05 pmin=-.05 pmax=1.05 linkmod=H linkunit=cm theta=0.5;
   run;

ods html body="MC.htm" frame="MCframe.htm" contents="MCcontents";
   proc print data=output3;
   title 'Parameter Estimates In Range';
   run;
ods html close;

proc btl data=marker.input2 map=marker.map output=output3;
   marker m12 m17;
   model surv=;
   parmest /cross=B gen=1 grid=.025 pmin=-.2 pmax=1.2 linkmod=H linkunit=cm theta=0.5
boot=1000 r=.025 .025;
   run;
```